EPiC Series in Computing

Volume 64, 2019, Pages 61–67

Proceedings of 28th International Conference
on Software Engineering and Data Engineering

Socio-Analyzer: A Sentiment Analysis Using Social Media Data

Ajay Bandi and Aziz Fellah

Northwest Missouri State University
School of Computer Science and Information Systems
Maryville, MO 64468 USA
ajay@nwmissouri.edu afellah@nwmissouri.edu

Abstract

The usage of social media is rapidly increasing day by day. The impact of societal changes is bending towards the peoples' opinions shared on social media. Twitter has received much attention because of its real-time nature. We investigate recent social changes in MeToo movement by developing Socio-Analyzer. We used our four-phase approach to implement Socio-Analyzer. A total of 393,869 static and stream data is collected from the data world website and analyzed using a classifier. The classifier identify and categorize the data into three categories (positive, neutral, and negative). Our results showed that the maximum peoples' opinion is neutral. The next higher number of peoples' opinion is contrary and compared the results with TextBlob. We validate the 765 tweets of weather data and generalize the results to MeToo data. The precision values of Socio-Analyzer and TextBlob are 70.74% and 72.92%, respectively, when considered neutral tweets as positive.

1 Introduction

A social issue is a problem that influences a considerable number of individuals within a society or around the world. It is often caused by several factors that arise from a conflict between diverse population. In other words, disputes' occur as a result of all forms of differences among individuals concerning the culture, gender, age, ability, religion, personality, social status, and sexual orientation. Such social issues become social movements and may lead to special law for the people at the workplace or in society. Few examples of social movements [10] are Black Lives Matter, Me Too, Ferguson, Same-Sex Marriage, Keep Families Together, etc. Other social changes include protecting human rights, animal welfare, banning plastic, etc. In today's world, social media play a crucial role in social movements [14]. People are open to post their feelings, opinions, reactions, and emotions towards the campaign [13]. Sometimes we can observe these movements become social media [5] protests or rallying calls. Sociologists study the people's responses on social movements either by conducting interviews, surveys or through opinion polls. However, this is a tedious process and limited to only a few samples. In this paper, we integrate the data science with sociology and analyze the vast data from social media by using machine learning algorithms [12]. We mined the Twitter data using hashtags (Metoo) and API calls and analyze the results.

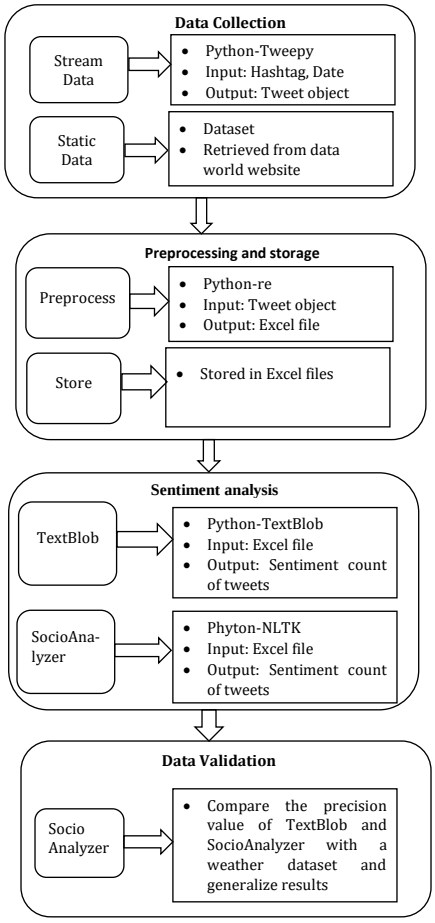

Figure 1: Methodology

The main goal of our research is to develop a Socio-Analyzer that analyzes the sentiment [6, 7] of social media data related to social movements. The rest of the paper is organized as follows. Section 2 presents the methodology of our data science project. Section 3 discusses the results and analysis of hashtag MeToo data. Finally, Section 4 presents conclusions and future work.

2 Methodology

Our workflow of collecting and analyzing tweets is in four different phases. They are 1) Data Collection 2)Preprocessing and Storage, 3)Sentiment Analysis, and 4)Validating the Results. These phases are shown in Figure Figure 1.

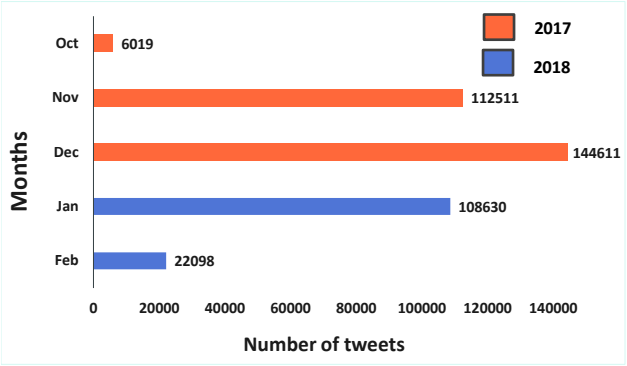

Figure 2: Data description

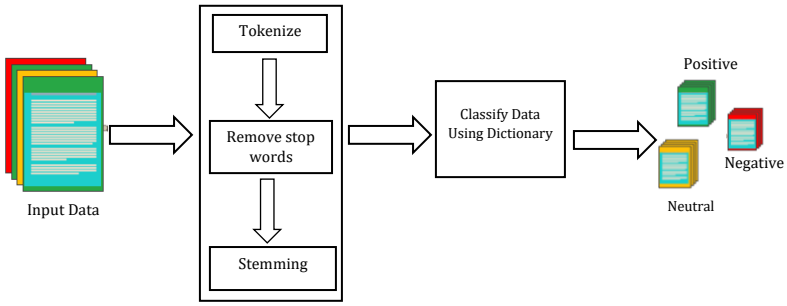

Figure 3: Sentiment Analysis

2.1 Data Collection

We divided our data into two types, static and stream data. Static data, hashtag MeToo data, is retrieved from the data world website. We scraped the stream data, hashtag MeToo data, periodically from the Twitter using Python-Tweepy library. The hashtag and the date are the inputs to the Tweepy library. The output is the tweet object for the given hashtag for one week concerning the input date. We collected the tweets from October 2017 to February 2018, which overlap with the tweets retrieved from the data world website. To eliminate the redundancy, we use only the data from the data world website for our analysis. The total number of tweets that we retrieved from the data world website on hashtag MeToo are 393,869. The length of each tweet is 140 characters that include the actual message, hashtag, emoticons, etc. Emoticons emphasis the emotion of the user [8, 15]. The description of data is shown in Figure 2.

2.2 Preprocessing and Storage

Preprocessing data is a fundamental step for any Natural Language Processing (NLP) [9]. We used the regular library in Python to remove the random patterns in the tweet. For instance, emoticons and embedded URLs and hashtags in the tweet. The cleaned tweets are stored in the Excel file. The cleaned data consists of tweetId, dateOfTweet, and text (no URLs or hashtags).

2.3 Sentiment Analysis

We developed our Socio-Analyzer to determine whether the people reacting on MeToo movement are supporting positively or negatively or neutral. The significant steps in developing SocioAnalyzer is building a sentiment dictionary [1, 3, 4, 9] using Natural Language Analyses with NLTK and classify the test data into one the three categories (positive, neutral, and negative) [2]. Figure 3 shows the process of analyzing the data and categorizing the data.

The input data is the set of tweets. Each tweet is then separate all the tokens considering space as a delimiter. Then, remove all the stopping words [11]. Stopping words are defined in the NLTK. The resultant set is consists of the multiple forms of the words. Count these words as one word using stemming. The approach to extract sentiment from tweets is as follows from the dictionary.

1. Start downloading and caching the sentiment dictionary
2. Download twitter testing data sets and input it into the program
   Input: It is beautiful day to go for fishing
3. Tokenize each word in the data set and feed in to the program
   ['it', 'is', 'beautiful', 'day', 'to', 'go', 'for', 'fishing']
4. Clean the tweets by removing the stop words
   ['beautiful', 'day', 'go', 'fishing']
5. The multiple forms of each word are counted as one word using stemming.
   ['beautiful', 'day', 'go', 'fishing']
   ['beauti', 'day', 'go', 'fish']
6. Now, for each word, compare it with positive and negative sentiments word in the dictionary. If matches, then increment positive count or negative count.

2.4 Data Validation

Data validation is an essential phase for any data science projects. However, validating 393,869 records of data is very challenging. It is tedious to verify all the tweets manually. We selected 765 tweets of weather data set to benchmark for our project. We manually categorized 765 into positive, neutral, and negative and verified the manual results using TextBlob and Socio-Analyzer. For validation, we compared the precision of TextBlob and our Socio-Analyzer. We calculated the precision considering two cases — case 1: Neutral values as positive and case 2: Neutral values as negative.

While there are several sentiment analysis tools available in the market, we selected TextBlob to validate our results. We conducted a pilot study for choosing the tool. We randomly selected ten random tweets and compared with three different tools (TextBlob, ParallelDots, Aylien). Of these three, TextBlob gives accurate results for all the ten tweets. The results are given in Table 1.

3 Results and Analysis

The primary goal of our research is to build the Socio-Analyzer to analyze the social media data using sentiment analysis. We analyzed tweets related to MeToo from Twitter to investigate how

Table 1: Results of Random Tweets

Text	Manual	TextBlob	ParallelDots	Aylien
love this sandwich	+ve	+ve	+ve	+ve
This is an amazing place!	+ve	+ve	+ve	+ve
I feel very good about these beers	+ve	+ve	+ve	+ve
This is my best work	+ve	+ve	+ve	+ve
What an awesome view?	+ve	+ve	+ve	+ve
Tomorrow is Wednesday	Neu	Neu	Neu	Neu
I am tired of this stuff	-ve	-ve	-ve	-ve
I can't deal with this	-ve	-ve	-ve	-ve
He is my sworn enemy!	Neu	Neu	-ve	+ve
I am here	Neu	Neu	Neu	Neu

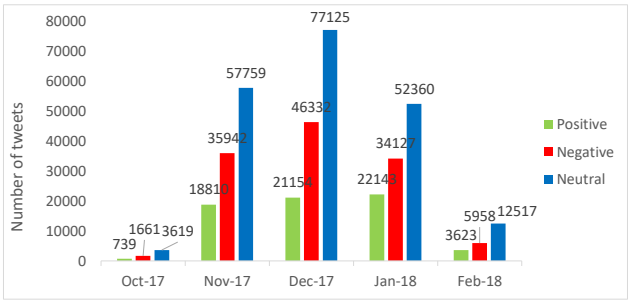

Figure 4: MeToo Dataset Results

people on social media react towards the social movement. This is the most recent active social movements. There are local and international alternative names for this movement, and the keywords associated with this are sexual harassment or sexual assault. The data collection process is explained in Section 2.1. These tweets have been collected over a period of 5 months from October 2017 to February 2018.

The results of the sentiment analysis of the collected tweets are shown in Figure 4. The results are categorized into positive, neutral, and negative categories. The positive effects are how people supportive of the cause (MeToo movement). Our results are consistent among the five different months. There are more people neither supportive nor opposed to the movement. Then, the second highest category of people is non-supportive to the MeToo movement. The last group of people who are supportive of the MeToo movement. These results are consistent among the five months of the data we analyzed.

We validate the weather data, as explained in Section 2.4. The results of data validation are shown in Figure 5. We manually evaluated the weather dataset then analyzed the data with Socio-Analyzer and compared the results with TextBlob.

We calculated the precision for the weather data using both the Socio-Analyzer and TextBlob. Precision is a measure of the variation among survey estimates is using the formula. $Precision = TruePositive/(TruePositive + FalsePositive)$. In our results, we have three categories, i.e., positive, negative, and neutral. For comparison purpose, we calculated two precision values considering neutral as a positive and neutral as a negative value. The values of Socio-Analyzer and the TextBlob's precision are given in Table 2. The precision values of Socio-Analyzer and TextBlob are 70.74% and 72.92%, respectively, when considered

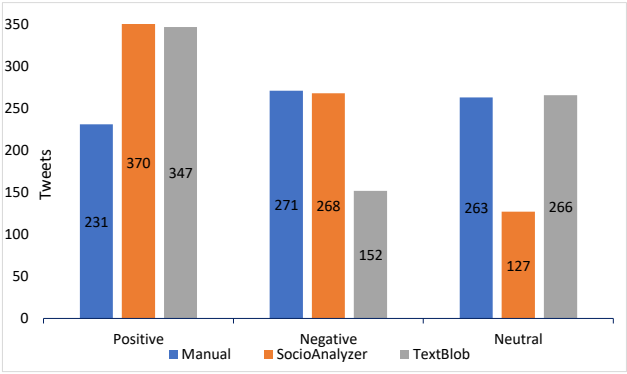

Figure 5: Data Validation Results

Table 2: Precision values

Tool	TP	FP	Precision
Considering neutral as positive			
Socio-Analyzer	428	177	0.707438
TextBLob	447	166	0.729201
Considering neutral as negative			
SocioAnalyzer	151	68	0.689498
TextBLob	183	95	0.658273

neutral tweets as positive. The precision values of Socio-Analyzer and TextBlob are 68.94% and 65.82%, respectively, when considered neutral tweets as negative.

4 Conclusions and Future Work

As described in the manuscript, we analyze the tweets of hashtag MeToo. The results show that the majority of the tweets (52%) are neutral (neither supportive nor oppose) and the next largest group's (31%) opinion is negative. Out of 393,869 tweets only 66,469 (17%) are positive. We compared our Socio-Analyzer results with the TextBlobs results, and the precision values of Socio-Analyzer and TextBlob are 70.74% and 72.92%, respectively, when considered neutral tweets as positive. This shows that our Socio-Analyzer analysis is closer to the TextBlob. The major limitations of our study are: unable to obtain the demographic information (age, geographic location, and gender) of users from the tweets. Collection of tweets can be possible only for the dates of one week. It is tedious and expensive to collect data for every week.

References

[1] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 36–44. Association for Computational Linguistics, 2010.

[2] Luiz Fernando Sommaggio Coletta, Nádia Félix Felipe da Silva, Eduardo Raul Hruschka, and Estevam Rafael Hruschka. Combining classification and clustering for tweet sentiment analysis. In *2014 Brazilian Conference on Intelligent Systems*, pages 210–215. IEEE, 2014.

[3] Miks Q Cureg, Juan Aurel D De La Cruz, Juan Carlos A Solomon, Aresh T Saharkhiz, Ariel Kelly D Balan, and Mary Jane C Samonte. Sentiment analysis on tweets with punctuations, emoticons, and negations. In *Proceedings of the 2019 2nd International Conference on Information Science and Systems*, pages 266–270. ACM, 2019.

[4] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[5] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*, 2011.

[6] Clement Levallois. Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 414–417, 2013.

[7] Quanzhi Li, Qiong Zhang, and Luo Si. Tweetsenti: Target-dependent tweet sentiment analysis. In *The World Wide Web Conference*, pages 3569–3573. ACM, 2019.

[8] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *Twenty-sixth aAAI conference on artificial intelligence*, 2012.

[9] Bryan Pratama, Dedi Dwi Saputra, Deny Novianti, Endah Putri Purnamasari, Antonius Yadi Kuntoro, Windu Gata, Nia K Wardhani, Sfenrianto Sfenrianto, Sularso Budilaksono, et al. Sentiment analysis of the indonesian police mobile brigade corps based on twitter posts using the svm and nb methods. In *Journal of Physics: Conference Series*, volume 1201, page 012038. IOP Publishing, 2019.

[10] Tushar Rao and Saket Srivastava. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)*, pages 119–123. IEEE Computer Society, 2012.

[11] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *International semantic web conference*, pages 508–524. Springer, 2012.

[12] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962. ACM, 2015.

[13] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1397–1405. ACM, 2011.

[14] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

[15] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1528–1531. ACM, 2012.